Springboard--Data Science Career Program

Capstone Project #2 - Yelp Sentiment Analysis

Final Report

By Kevin Cole

August 2020

1. **Introduction**

   Positive yelp reviews are crucial for a business. Yelp not only drives traffic to business but easily helps users make decisions about which establishment to visit. These decisions are often made based on reviews and ratings found on the yelp platform.

   A restaurant owner may want to tailor the customer experience to items which lead to high reviews and make business decisions to eliminate/reduce anything that can lead to negative reviews.

   The goal is to use data science methods to establish a connection between the rating levels and what might drive them.

2. **Approach**

   2.1 Data Acquisition and Wrangling

   The data is provided by Yelp as an open dataset (https://www.yelp.com/dataset). This subset of Yelp data is a 10.5 GB file containing 5 different JSON encoded files. These files contain data on the business, the customer review, the user, the datetime of visit, and additional tips left by the user.

   After exploration of the files, it was determined that for this project the information in the business data and the customer review data was the data of interest. 'yelp_academic_dataset_business.json' contains business info for the business in the review set. Each business is associated with some categories, information about being open/closed, hours of operation, and a unique business ID to tie the business with the appropriate reviews.

   'yelp_academic_dataset_review.json' contains reviews of a large variety of businesses, their textual reviews, and the stars given by the user. This file is quite large (6.33 GB).

   The business information contains a categories variable that describes all categories of business type that the business falls into. After viewing the categories in the data, the “Restaurant” category was determined to be the most valuable category tag as opposed to a more specific category like “Asian”, “Mexican”, or “Pizza”. The business information dataset was then subsetted to include only businesses that fall into the “Restaurant” category, removing other non-food centric businesses.

The reviews dataset was then merged with the restaurant data on the unique business ID present in both datasets. Reviews for businesses not in the restaurant data were removed.

The variables of interest in the resulting dataframe were the number of stars the review gave, the review text, and a flag for the operating status of the restaurant for open/closed subsetting.

A good/neutral/bad categorical variable was created mapping 4 and 5 star ratings to good, 1 and 2 star ratings to bad, and 3 to neutral. The resulting dataset contained slightly more than 5 million Yelp reviews on restaurants. This is down from the 8 million reviews in the dataset of all reviews.

The text reviews still needed to be cleaned. The text cleaning involved the following steps.

- Reviews shorter than 11 characters were deemed not valuable as the seriousness of their review is in question. These reviews were removed.
- Review text was converted to all lowercase.
- Special characters including punctuation and carriage returns were removed.
- Text language was detected using the “langdetect” python module and non-english reviews were removed. This was a computationally time intensive step.
- Stopwords defined in the NLTK python library were removed.
- Words with a frequency of less than 20 in the dataset were removed. This accounted for odd text, typos, and other non-english tokens. With 5,026,166 reviews in the dataset, words not appearing in at least 20 documents (0.0004%) are very likely not influential words.

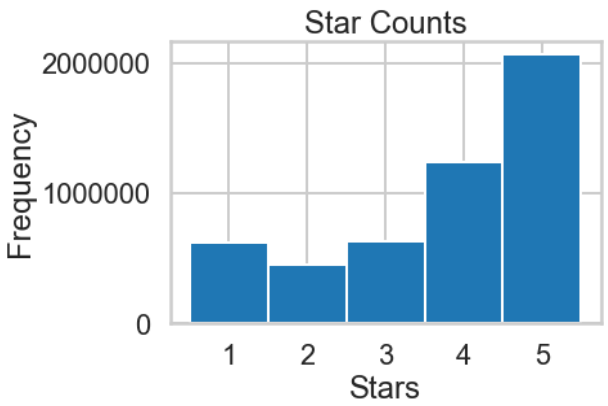
The final cleaned dataset has 5,026,166 reviews and the dataframe was saved off using the pickle format. This cleaning reduced the size of the original dataset from 10.5 GB to 1.97 GB.

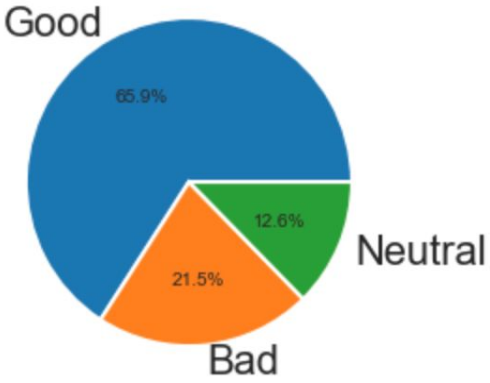
	stars	text	good_bad	language	is_open
0	5	love deagans really atmosphere cozy festive sh...	Good	en	1.0
1	1	dismal lukewarm defrostedtasting texmex glop m...	Bad	en	0.0
2	4	oh happy day finally canes near casa yes other...	Good	en	1.0
3	5	definitely favorite fast food sub shop ingredi...	Good	en	1.0
4	5	really good place simple decor amazing food gr...	Good	en	1.0
...	...	...	...	...	...
5026161	5	confections cash casinos welcome las vegas fin...	Good	en	0.0
5026162	3	solid american food southern comfort flare war...	Neutral	en	1.0
5026163	5	im honestly sure never place im definitely goi...	Good	en	1.0
5026164	3	food decent say service took way long order ev...	Neutral	en	1.0
5026165	5	oh yeah service good food good serving good se...	Good	en	1.0

5026166 rows × 5 columns

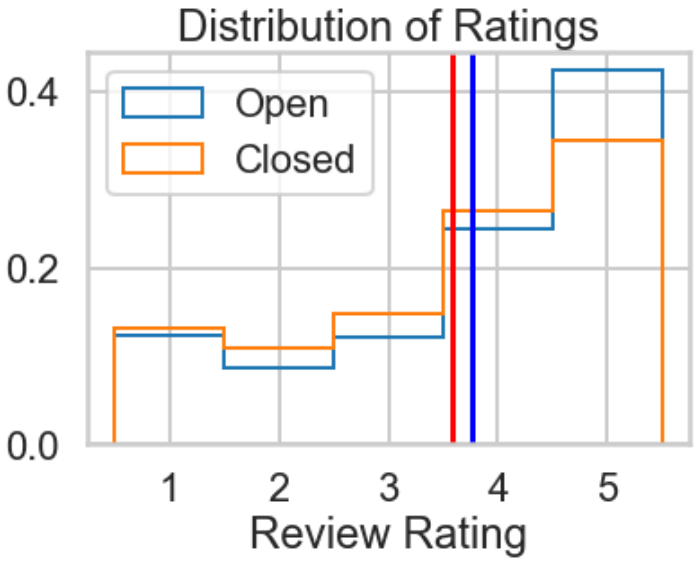
2.2 Exploratory Data Analysis:

The distribution of star counts is not normal, it is skewed towards higher ratings. This is likely due to the tendency of people to rate interactions as good unless they have a specific reason not to. The default rating tends to be 4-star or 5-star if a person has a reasonably normal interaction which produces a review set of 2/3rds of reviews classified as good reviews.





This distribution of star ratings is similar between open and closed restaurants with a noticeable drop in percentage of 5-star reviews. Restaurants receiving 5-star reviews would be less likely to close, although there are certainly other reasons for restaurants closing.

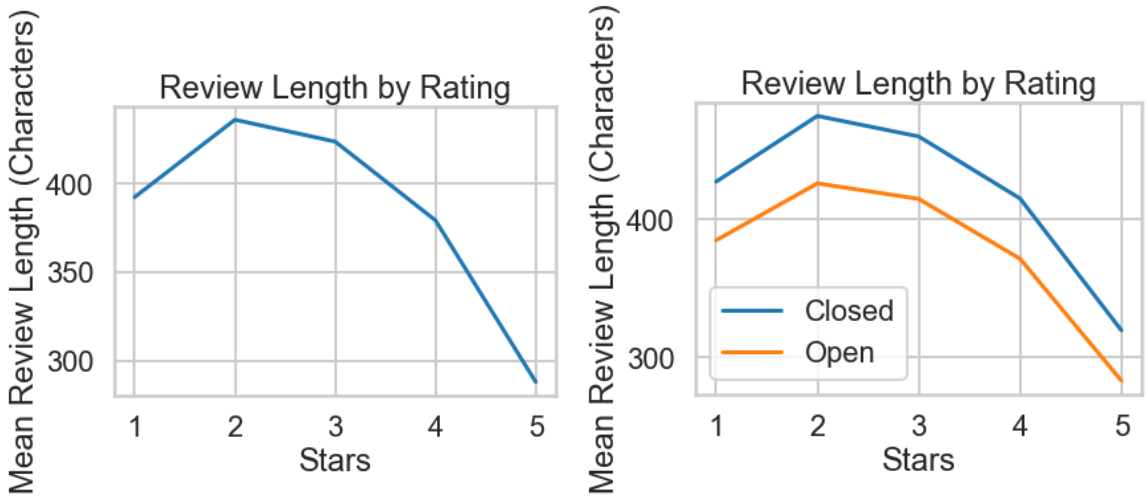


Mean rating for open restaurants: 3.76
Mean rating for closed restaurants: 3.57

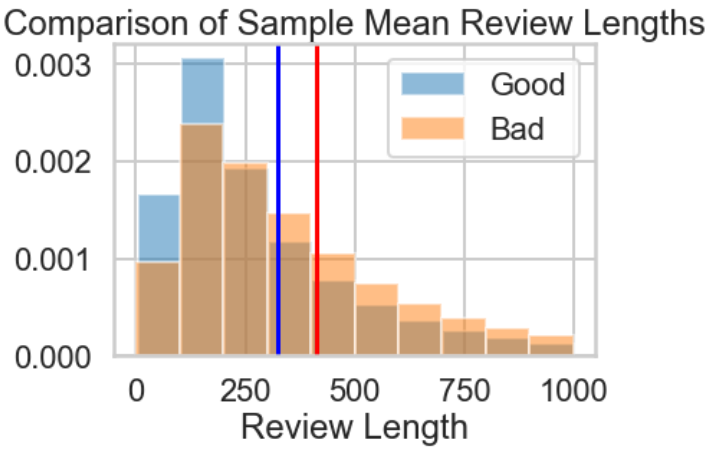
Difference observed: 0.18

The two samples were used to create a bootstrapped replicate array of difference in means. In the set of replicates there was not one single occurrence in 10,000 samples of a difference in means as large as the one observed. We can conclude that this difference in the mean rating of open restaurants and closed restaurants is statistically significant.

Another key indicator to the strength of the rating is the length of the review. When people have a poor experience, they want to talk about it, and share their experience. They want to warn future patrons about what they experienced and want to elaborate, sometimes turning their review into a place to rant. For a good experience, they tend to mention a few things that they liked but do not tend to drone on and on about the experience. This tendency can be seen in the reviews of both open and closed restaurants.

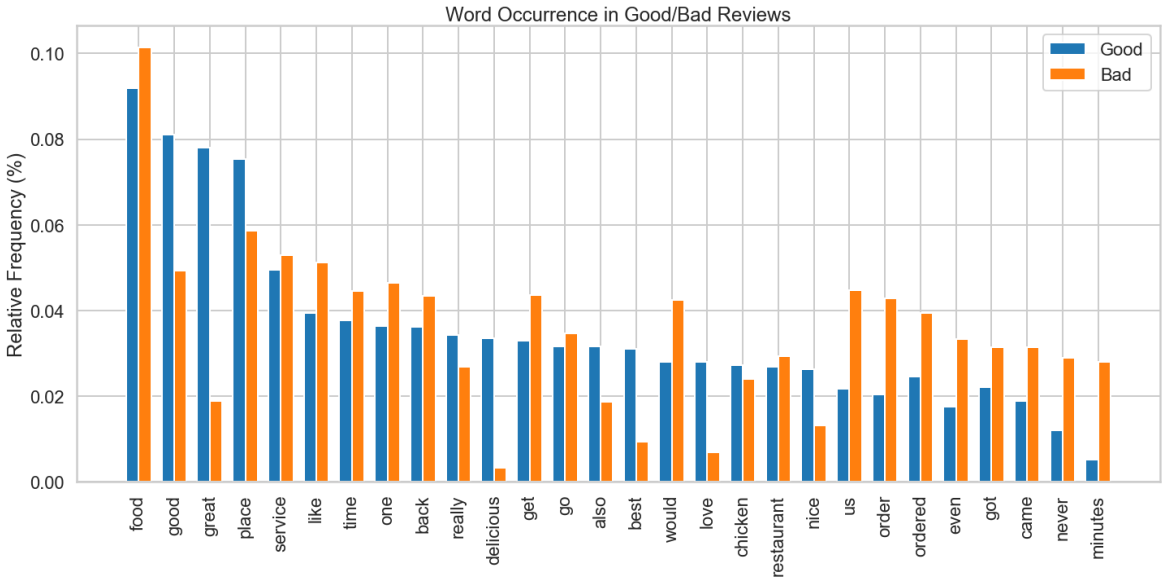


When examining the review length between positive and negative reviews, we can see a difference in the means that has been determined to be a statistically significant difference.



The frequency of words used is a big part of natural language processing. The types of words used will play a key role in determining the strength of the review. When looking at the word frequency in both good and bad reviews we can see some predictable words and some interesting differences. Good reviews tend to be polished by strong adjectives like “good”, “great”, and “delicious”. These strong adjectives are not only prevalent in good reviews but very unlikely to occur in a bad review.

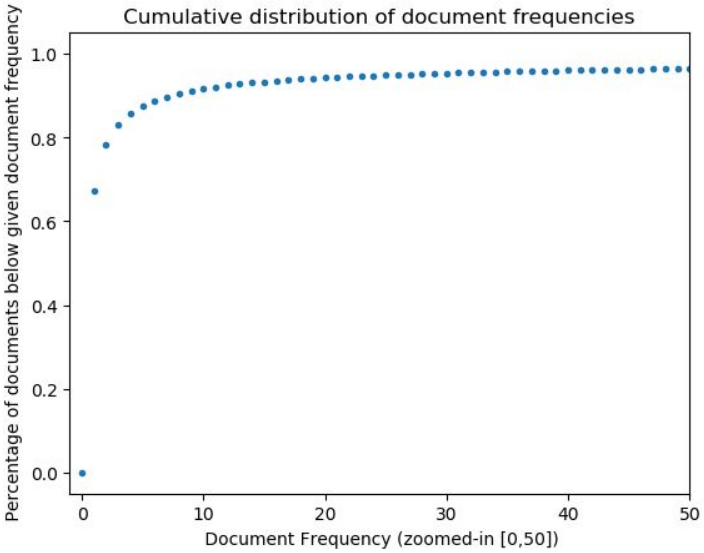
Words in negative reviews tend to be centered around service. Words like order/ordered, got/get, came, never, minutes dominate these reviews. This seems to be the largest point of complaints.



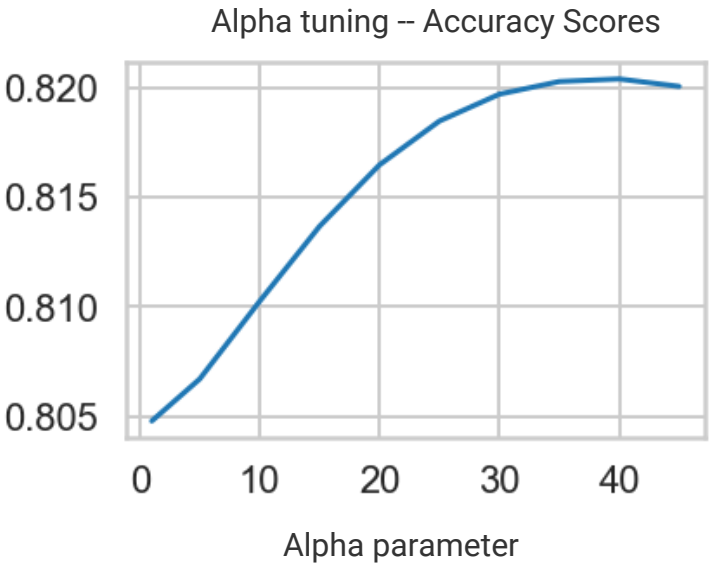
This can be seen in the word clouds created from word frequency as well.

2.3 Baseline Modeling

Baseline classification models were evaluated on percent accuracy, recall, and F1 score. Naive Bayes, Logistic Regression and Random Forests were all explored. The document frequency minimum was set at 20, removing words that appear in less than 0.0004% of the five million reviews.



The Naive Bayes model requires tuning of the alpha hyperparameter. A grid search was conducted to find the appropriate smoothing parameter. An alpha of 40 produced the best scores.



We can see from the classification report that while the model performs well on positive reviews (class 2), it has less strong performance on negative reviews (class 1), and poor performance on the neutral reviews.

```
Naive Bayes Test Report:
              precision    recall  f1-score   support

           0       0.76      0.77      0.76    216066
           1       0.44      0.36      0.40    127024
           2       0.90      0.92      0.91    662144

    accuracy                           0.82   1005234
   macro avg       0.70      0.68      0.69   1005234
weighted avg       0.81      0.82      0.81   1005234


Naive Bayes Train Report:
              precision    recall  f1-score   support

           0       0.76      0.77      0.76    864264
           1       0.44      0.36      0.40    508093
           2       0.90      0.92      0.91   2648575

    accuracy                           0.82   4020932
   macro avg       0.70      0.69      0.69   4020932
weighted avg       0.81      0.82      0.81   4020932
```

Logistic Regression performed slightly better overall. The improvement was in the negative and neutral classes. The initial Logistic Regression model showed almost no overfitting, so tuning of the model was not required. Additionally, investigations of the regularization parameter C showed similar performance across many values.

```
Logistic Regression Test Report:
              precision    recall  f1-score   support

           0       0.82      0.84      0.83    216066
           1       0.56      0.31      0.40    127024
           2       0.89      0.96      0.93    662144

    accuracy                           0.85   1005234
   macro avg       0.76      0.70      0.72   1005234
weighted avg       0.84      0.85      0.84   1005234


Logistic Regression Train Report:
              precision    recall  f1-score   support

           0       0.82      0.84      0.83    864264
           1       0.58      0.32      0.41    508093
           2       0.89      0.96      0.93   2648575

    accuracy                           0.86   4020932
   macro avg       0.76      0.71      0.72   4020932
weighted avg       0.84      0.86      0.84   4020932
```

The Random Forest model required some tuning as with a dataset this large, building the trees out too large is computationally intensive and time consuming. As a result, the maximum features used in a tree was limited by the log (base 2) of the number of datapoints. Anything larger became too intensive to consider. Max depth of each tree was set to 50, and the number of trees in the forest was tuned to 1000. None of the classes performed better than the logistic regression model.

Random Forest Test Report

	precision	recall	f1-score	support
0	0.76	0.68	0.72	216066
1	0.34	0.38	0.36	127024
2	0.85	0.86	0.85	662144
accuracy			0.76	1005234
macro avg	0.65	0.64	0.64	1005234
weighted avg	0.77	0.76	0.76	1005234

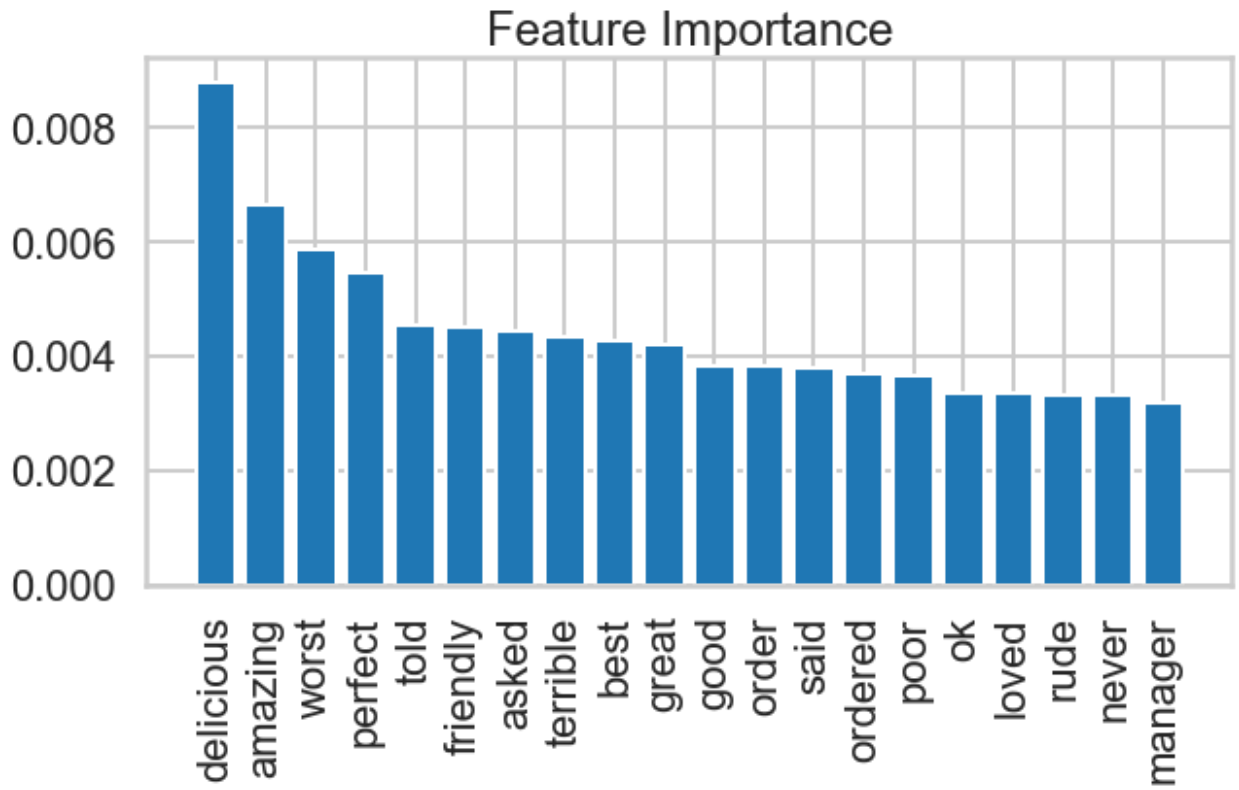
Random Forest Train Report:

	precision	recall	f1-score	support
0	0.79	0.70	0.74	864264
1	0.42	0.44	0.43	508093
2	0.86	0.88	0.87	2648575
accuracy			0.79	4020932
macro avg	0.69	0.68	0.68	4020932
weighted avg	0.79	0.79	0.79	4020932

2.4. Feature Importance

Feature frequency in reviews as reviewed above gives us some insight as to how often words are used in positive ratings as well as negative ratings, but does not give us the entire picture. For example, “food” was highly prevalent in both classes mainly because the reviews are about restaurants and not because the mention of the word food is inherently positive or negative.

Feature importance scores from the random forest model give us a great sense of the words that most directly influence the sentiment of the review. We see some of the same themes as the word frequency analysis, but words like “food” that appear in both positive and negative reviews are not prevalent. This analysis, while telling us which words have strong influence, does not tell us which class the token is pushing the review into.



The logistic regression model is particularly useful for this as the coefficients of each feature in the decision function can be multiplied by the standard deviation of the corresponding parameter data to create a standardized score for the influence of each token on each class.

Examining these results in a word cloud gives us a different view of feature importance than word counts alone did.

We can see that there are several words that did not show up prominently in the frequency count analysis. This offers us a set of words that are used not just because they are in food reviews, but specifically tuned to the sentiment of the review. Strong adjectives dominate these clouds as reviewers only use words like “wonderful”, “yummy”, and “superb” when describing a negative experience. These words could be talking about many different facets of their experience, but are all positive in nature.

2.5. Extended Modeling

2.5.1 Latent Dirichlet Allocation

A Latent Dirichlet Allocation (LDA) model was used as an unsupervised way to group similar parts of the data into topics. As these are not predefined topics, they are unnamed but we can get a sense for the positive/negative nature of each category based on the tokens that show up within that category.

```
Topics found via LDA:

Topic #0:
food ice place cream like tea service menu better price

Topic #1:
pizza burger good fries like cheese sandwich place salad ordered

Topic #2:
great place food good staff friendly nice bar service love

Topic #3:
good food great place really service steak delicious like dinner

Topic #4:
good tacos fish ordered food really great sushi like chicken

Topic #5:
chicken sauce dish restaurant delicious pasta ordered bread menu dishes

Topic #6:
happy hour like night coffee bar nice place best great

Topic #7:
food order service time minutes came table asked said got

Topic #8:
food service place great good time breakfast best ive vegas

Topic #9:
food good place chicken rice like soup thai restaurant noodles
```

Most topics are positive in nature which is not entirely a surprise as 2/3rds of data was positive in nature. Topics tended to separate by restaurant type, often bringing in not just the positive adjectives we saw in feature importance but the type of food that is being described as well.

Topic #7 is the notable topic as while not explicitly negative, this is the topic in which most negative reviews might fall as these words are service related and there are no positive words in the topic. These reviews tend to be talking about the negative points of the experience, such as wait time and service quality.

2.5.2. Bigram tokens

Expanding the vectorizer to allow for bigrams allows for the meaning of words to combine to create new or stronger sentiment. Many adjectives in particular rely on the words that follow to ascertain the positive/negative nature of their meanings.

When allowing the vectorizer to introduce bigrams to the corpus, the number of tokens jumps from 74k to 980k, since many new combinations of word pairs are being considered.

As far as performance, the logistic regression model still outperformed other models considered in the baseline analysis, and improved on its own performance slightly.

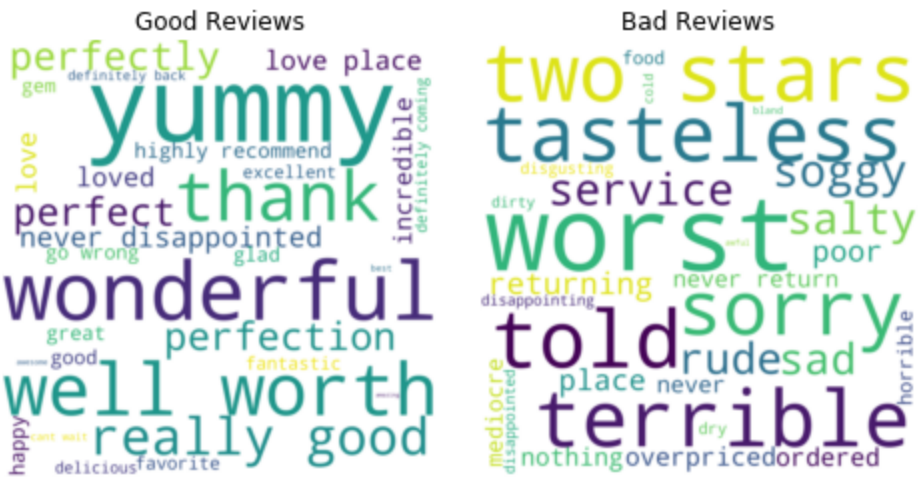
Logistic Regression Test Report:

	precision	recall	f1-score	support
0	0.84	0.86	0.85	216066
1	0.59	0.40	0.48	127024
2	0.91	0.96	0.94	662144
accuracy			0.87	1005234
macro avg	0.78	0.74	0.76	1005234
weighted avg	0.86	0.87	0.86	1005234

Logistic Regression Train Report:

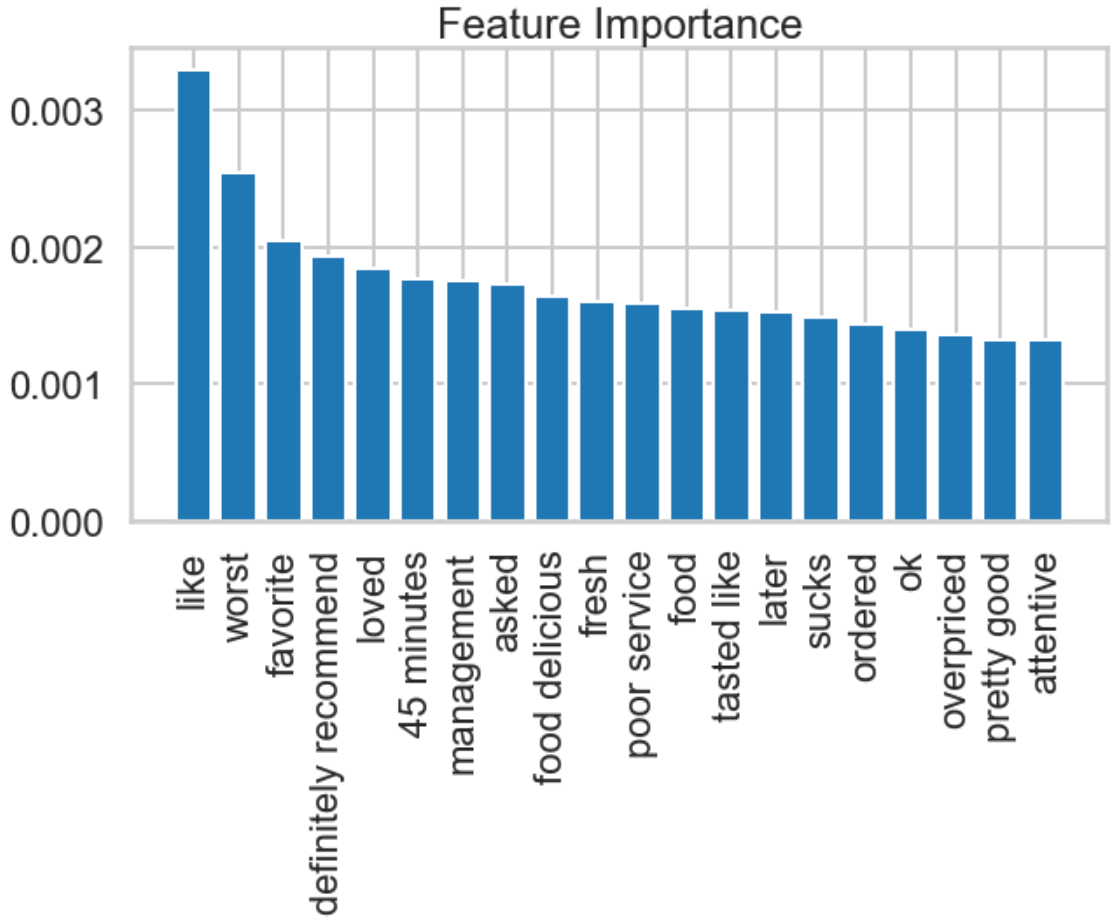
	precision	recall	f1-score	support
0	0.86	0.88	0.87	864264
1	0.66	0.47	0.55	508093
2	0.92	0.97	0.94	2648575
accuracy			0.89	4020932
macro avg	0.81	0.77	0.79	4020932
weighted avg	0.88	0.89	0.88	4020932

Introductions of the bigrams had some influence on the most influential tokens for both positive and negative reviews. In the word clouds, you can see the presence of strong pairings that are more meaningful when together like “love place” and “never return” . The word “return” alone could be in a positive or negative review but next to the word “never” provides a decisively negative context.



Word clouds including bigram tokens

We can also see in the feature importance graph some bigrams emerge. Interestingly as far as service goes, 45 minutes seems to be a threshold where customers become dissatisfied with the service.



3. Conclusions and Future Work

The model that best predicted the sentiment of a review was the Logistic Regression Model. While all models had minimal overfitting, the Logistic model produced the highest accuracy, both overall and within each class.

The neutral class was the biggest problem in each model. This is due to the small nature of the class, as well as the overlap of both positive and negative words appearing in such reviews.

Future work could include:

- Examine effects of grouping neutral reviews with negative reviews for a two class problem. This would reduce the effects of having drastically different class sizes.
- Cleaning of the dataset could use more work. Even with a min_df of 20 (removing words that did not occur in a minimum of 20 documents) there were still nonsensical words and non-english tokens in the dataset. This work would include exploring more robust vectorizers.
- Examine the effects of increasing token length to more than two words.
- Topics found in LDA could be used for subsetting analysis.
- Analysis could be tailored towards particular restaurant types. For example, analysis could be conducted on specifically pizza parlor reviews if a client was interested in insights specific to that topic.
- Deeper trees and larger forests in the random forest model. Current analysis was limited by hardware and time constraints. With time and computing power, this model could be explored more.

4. Recommendations for client

There were two main trends that drove the sentiment of a review, service quality and food quality. The first and largest indicator of a negative review is around service. When the customer waits too long, has issues with their order, or has a poor experience with the server, their experience is soured enough that it dominates their review, regardless of food quality. Owners and managers should create environments and staff expectations that minimize these poor interactions, as they are the fastest way to bad reviews.

In addition, these service related challenges tended to cause the rant phenomenon. Reviews with longer length tended towards the negative ratings as people who do not have a good experience like to rant about it. They will go into detail and keep typing. When customers have a good experience, they mention a few items of note and move on. Avoiding these service related issues is the best way to keep the ranting reviewer at bay.

Once service related challenges have been addressed, likely through training and procedure, food related challenges can be considered next most important. The word food is the most prevalent word across all reviews -- people want to talk about their food particularly if they aren't specifically complaining about an experience that they wish to warn others about (like the 30,000 bad reviews that mention the word "hair"). While the food quality is largely in the hands of the chef but a business owner should be cognisant of the fact that for a top-tier review, adjectives that express true delightment will come out more frequently with high-quality food.

5. Consulted Resources

- Yelp - initial dataset - https://www.yelp.com/dataset
- Data Camp -- Python skills and models influenced by data camp exercises. List of courses used can be seen at https://www.datacamp.com/profile/kevinccole
- CS109 lectures provided via Springboard curriculum.
- Documentation for imported packages.
- Stack Overflow for individual coding questions.
- Consultations with Data Science professional AJ Sanchez.